Using Subcategorization to Resolve Verb Class Ambiguity

Using Subcategorization to Resolve Verb Class Ambiguity

Maria Lapata
School of Cognitive Science
Division of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
mlap@cogsci.ed.ac.uk

Chris Brew
HCRC Language Technology Group
Division of Informatics
University of Edinburgh
2 Buccleuch Place
Edinburgh EH8 9LW, UK
chrisbr@cogsci.ed.ac.uk

Abstract

Levin's (1993) taxonomy of verbs and their classes is a widely used resource for lexical semantics. In her framework, some verbs, such as *give* exhibit no class ambiguity. But other verbs, such as *write*, can inhabit more than one class. In some of these ambiguous cases the appropriate class for a particular token of a verb is immediately obvious from inspection of the surrounding context. In others it is not, and an application which wants to recover this information will be forced to rely on some more or less elaborate process of inference. We present a simple statistical model of verb class ambiguity and show how it can be used to carry out such inference.

1 Introduction

The relation between the syntactic realization of a verb's arguments and its meaning has been extensively studied in Levin (1993). Levin's work relies on the hypothesis that "the behavior of a verb, particularly with respect to the expression and interpretation of its arguments, is to a large extent determined by its meaning" (Levin, 1993, p. 1). Verbs which display the same *diathesis alternations*–alternations in the realization of their argument structure–are assumed to share certain meaning components and are organized into a semantically coherent class.

As an example consider sentences (1)–(3) taken from Levin. Example (1) illustrates the causative/inchoative alternation. Verbs undergoing this alternation can be manifested either as transitive with a causative reading (cf. (1a)) or as intransitive with an inchoative reading (cf. (1b)). Examples (2) and (3) illustrate the dative and benefactive alternations respectively. Verbs which license the former alternate between the prepositional frame NP-V-NP-PP_{to} (cf. (2a)) and the double object frame V-NP-NP (cf. (2b)), whereas verbs which undergo the latter alternate between the double object frame (cf. (3a)) and the prepositional frame NP-V-NP-PP_{for} (cf. (3b)).

(1) a. Janet broke the cup.
    b. The cup broke.
(2) a. Bill sold a car to Tom.
    b. Bill sold Tom a car.
(3) a. Martha carved the baby a toy.
    b. Martha carved a toy for the baby.

Verbs like *crack* and *chip* pattern with *break* in licensing the causative/inchoative alternation and are associated with the semantic class of BREAK verbs. Verbs *make* and *build* behave similar to *carve* in licensing the benefactive alternation and are members of the class of BUILD verbs, whereas *sell* and *give* undergo the dative alternation and participate in the GIVE class. By grouping together verbs which pattern together with respect to diathesis alternations Levin defines approximately 200 verb classes, which she argues reflect important semantic regularities.

2 Motivation

Levin provides an index of 3,024 verbs for which she lists the semantic classes and diathesis alternations. The mapping between verbs and classes is not one-to-one. Of the 3,024 verbs which she covers, 784 are listed as having more than one class. Even though Levin's monosemous verbs outnumber her polysemous verbs by a factor of nearly four to one, the total frequency of the former (4,252,715) is comparable to the total frequency of the latter (3,986,014). This means that close to half of the cases processed by a hypothetical semantic tagger would manifest some degree of ambiguity. The frequencies are detailed in table 1 and were compiled from a lemmatized version of the British National Corpus (BNC), a widely distributed 100 million word collection of samples of written and spoken English (Burnard, 1995).

Classes	Verbs	BNC frequency
1	2,239	4,252,715
2	536	2,325,982
3	173	738,854
4	43	395,212
5	23	222,747
6	7	272,669
7	2	26,123
10	1	4,427

Table 1: Polysemous verbs according to Levin

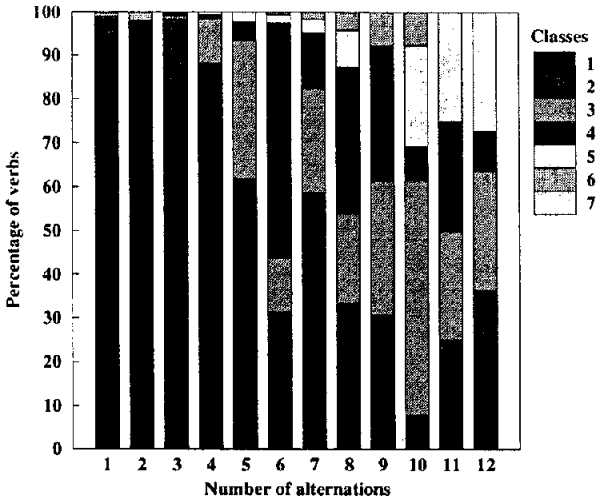

Figure 1: Relation between number of classes and alternations

Furthermore, as shown in figure 1, the number of alternations licensed by a given verb increases with the number of classes it inhabits. Consider for example verbs participating in one alternation only: of these, 90.4% have one semantic class, 8.6% have two classes, 0.7% have three classes and 0.3% have four classes. In contrast, of the verbs licensing six different alternations, 14% have one class, 17% have two classes, 12.4% have three classes, 53.6% have four classes, 2% have six classes and 1% has seven classes.

Palmer (1999) and Dang et al. (1998) argue that the use of syntactic frames and verb classes can simplify the definition of different verb senses. Beyond this, we claim that information about the argument structure of a polysemous verb can often help disambiguating it.

Consider for instance the verb *serve* which is a member of four classes: GIVE, FIT, MASQUERADE and FULFILLING. Each of these classes can in turn license four distinct syntactic frames. As shown in the examples[1] below, in (4a) *serve* appears ditransitively and belongs to the semantic class of GIVE verbs, in (4b) it occurs transitively and is a member of the class of FIT verbs, in (4c) it takes the predicative complement *as minister of the interior* and is a member of MASQUERADE verbs. Finally, in sentence (4d) *serve* is a FULFILLING verb and takes two complements, a noun phrase (*an apprenticeship*) and a prepositional phrase headed by *to*. In the case of verbs like *serve* we can guess their semantic class solely on the basis of the frame with which they appear.

(4) a. I'm desperately trying to find a venue for the reception which can serve our guests an authentic Italian meal.
b. The airline serves 164 destinations in over 75 countries.
c. Jean-Antoine Chaptal was a brilliant chemist and technocrat who served Napoleon as minister of the interior from 1800 to 1805.
d. Before her brief exposure to pop stardom, she served an apprenticeship to a still-life photographer.

But sometimes we do not have the syntactic information that would provide cues for semantic disambiguation. Consider sentence (5a). The verb *write* is a member of three Levin classes, two of which (MESSAGE TRANSFER, PERFORMANCE) take the ditransitive frame NP-V-NP-NP. In this case we have the choice between the "message transfer" reading (cf. (5a)) and the "performance" reading (cf. (5b)). This is an instance of the common problem of inferring the value of a hidden variable (in this case the "true class" of a particular instance of *write*). The same situation arises with the verb *phone* which is listed as a GET verb and an INSTRUMENT OF COMMUNICATION verb and in both cases can take the frame NP-V-NP-NP. In sentence (5c) the preferred reading is that of "get" instead of "instrument of communication" (cf. sentence (5d)).

(5) a. A solicitor wrote him a letter at the airport.
b. I want you to write me a screenplay called "The Trip".

[1] Unless stated otherwise the example sentences were taken from the BNC and simplified for clarification purposes.

c. I'll phone you a taxi.
d. As I entered the room I wished I'd thought of phoning a desperate SOS to James.

The objective of this paper is to address the verb class disambiguation problem by developing a probabilistic framework which combines linguistic knowledge (i.e., Levin's classification) and frame frequencies acquired from the BNC. Our initial experiments focus on the syntactic frames characteristic for the dative and benefactive alternations (cf. examples (2) and (3)). These frames are licensed by a fairly large number of classes: 19 classes license the double object frame, 22 license the NP-V-NP-PP_{to} frame and 14 classes license the NP-V-NP PP_{for} frame. The semantic and syntactic properties of these alternations have been extensively studied and are well understood (see Levin (1993) and the references therein). Furthermore, they are fairly productive and one would expect them to be well represented in a large corpus.

In section 3 we describe the statistical model and the estimation of the various model parameters, section 4 presents some preliminary results and section 5 contains some discussion and concluding remarks.

3 The Model

We view the choice of a class for a polysemous verb in a given frame as the joint probability $P(verb, frame, class)$ which we rewrite using the chain rule in (6).

$$(6)\quad P(verb, frame, class) = P(verb)\, P(frame|verb)\, P(class|verb, frame)$$

We also make the following independence assumption:

$$(7)\quad P(class|verb, frame) \approx P(class|frame)$$

The independence assumption reflects Levin's hypothesis that the argument structure of a given verb is a direct reflection of its meaning. Accordingly we assume that the semantic class determines the argument structure of its members without making reference to the individual verbs. By applying Bayes Law we write $P(class|frame)$ as:

$$(8)\quad P(class|frame) = \frac{P(frame|class)\, P(class)}{P(frame)}$$

By substituting (7) and (8) into (6), $P(verb, class, frame)$ can be written as:

$$(9)\quad P(verb, frame, class) \approx \frac{P(verb)\, P(frame|verb)\, P(frame|class)\, P(class)}{P(frame)}$$

We estimate the probabilities $P(verb)$, $P(frame|verb)$, $P(frame|class)$ and $P(class)$ as follows:

$$(10)\quad P(verb) \approx \frac{f(verb)}{\sum_i f(verb_i)}$$

$$(11)\quad P(frame|verb) \approx \frac{f(verb, frame)}{\sum_i f(verb, frame_i)}$$

$$(12)\quad P(frame|class) \approx \frac{f(class, frame)}{\sum_i f(class, frame_i)}$$

$$(13)\quad P(class) \approx \frac{f(class)}{\sum_i f(class_i)}$$

$$(14)\quad P(frame) \approx \frac{f(frame)}{\sum_i f(frame_i)}$$

It is easy to obtain $f(verb)$ from the lemmatized BNC. For the experiments reported here, syntactic frames for the dative and benefactive alternations were automatically extracted from the BNC using Gsearch (Keller et al., 1999), a tool which facilitates search of arbitrary POS-tagged corpora for shallow syntactic patterns based on a user-specified context-free grammar and a syntactic query. The acquisition and filtering process is detailed in Lapata (1999). We rely on Gsearch to provide moderately accurate information about verb frames in the same way that Hindle and Rooth (1993) relied on Fidditch to provide moderately accurate information about syntactic structure, and Ratnaparkhi (1998) relied on simple heuristics defined over part-of-speech tags to deliver information nearly as useful as that provided by Fidditch. We estimated $f(verb, frame)$ as the number of times a verb co-occurred with a particular frame in the corpus.

We cannot read off $P(frame|class)$ from the corpus, because it is not annotated with verb classes. Nevertheless we can use the information listed in Levin with respect to the syntactic frames exhibited by the verbs of a given class. For each class

Class	Frames
MANNER	NP-V-NP-PP$_{from}$, NP-V-NP, NP-V-PP$_{at}$, NP-V-NP-PRED
ACCOMPANY	NP-V-NP, NP-V-NP-PP$_{to}$
THROW	NP-V-NP-NP, NP-V-NP-PP$_{loc}$, NP-V-NP-PP$_{from}$-PP$_{to}$, NP-V-NP, NP-V-NP-PP$_{to}$, NP-V-NP-PP$_{at}$,
PERFORMANCE	NP-V, NP-V-NP, NP-V-NP-NP, NP-V-NP-PP$_{to}$, NP-V-NP-PP$_{for}$, NP-V-NP
GIVE	NP-V-NP-PP$_{to}$, NP-V-NP-NP
CONTRIBUTE	NP-V-NP-PP$_{to}$

Table 2: Sample of verb classes and their syntactic frames

we recorded the syntactic frames it licenses (cf. table 2). Levin's description of the argument structure of various verbs goes beyond the simple listing of their subcategorization. Useful information is provided about the thematic roles of verbal arguments and their interpretation. Consider the examples in (15): in (15a) the verb *present* is a member of the FULFILLING class and its theme is expressed by the prepositional phrase *with an award*, in (15b) the PP headed by *with* receives a locative interpretation and the verb *load* inhabits the SPRAY/LOAD class, whereas in (15c) the prepositional phrase is instrumental and *hit* inhabits the HIT class. None of the information concerning thematic roles was retained. All three classes (FULFILLING, SPRAY/LOAD and HIT) were assigned the frame NP-V-NP-PP$_{with}$'.

(15) a. John presented the student with an award.
b. John loaded the truck with bricks.
c. John hit the wall with a hammer.

Because we didn't have corpus counts for the quantity $f(class, frame)$ we simply assumed that all frames for a given class are equally likely. This means, for instance, that the estimate for P(NP-V-NP-NP$_{to}$|GIVE) is $\frac{1}{2}$ and similarly the estimate for P(NP-V|PERFORMANCE) is $\frac{1}{6}$ (cf. table 2). This is clearly a simplification, since one would expect $f(class, frame)$ to be different for different corpora, and to vary with respect to class size and the frequency of class members.

In order to estimate $P(class)$ we first estimate $f(class)$ which we rewrite as follows:

$$(16)\ f(class) = \sum_i f(verb_i, class)$$

Class	$size(class)$	$p(class\|amb\_class)$	$f(verb, class)$
THROW	27	0.40	7783.6
SEND	20	0.27	5253.9
GIVE	15	0.20	3891.8
MARRY	10	0.13	2529.6

Table 3: Estimation of $f(verb, class)$ for the verb *pass*

The estimate of $f(verb, class)$ for monosemous verbs reduces to the count of the verb in the corpus. Once again we cannot estimate $f(verb, class)$ for polysemous verbs directly. All we have is the overall frequency of a given verb in the BNC and the number of classes it is a member of according to Levin. We rewrite $f(verb, class)$ as:

$$(17)\ f(verb, class) = f(verb)p(class|verb)$$

We approximate $p(class|verb)$ by collapsing across all verbs that have the appropriate pattern of ambiguity:

$$(18)\ f(verb, class) \approx f(verb)p(class|amb\_class)$$

Here *amb_class*, the ambiguity class of a verb, is the set of classes that it might inhabit.[2] We collapse verbs into ambiguity classes in order to reduce the number of parameters which must be estimated: we certainly lose information, but the approximation makes it easier to get reliable estimates from limited data. In future work we plan to use the EM algorithm (Dempster et al., 1977) to uncover the hidden *class*, but for the present study, we simply approximate $p(class|amb\_class)$ using a heuristic based on class size:

$$(19)\ p(class|amb\_class) \approx \frac{size(class)}{\sum_{c \in amb\_class} size(c)}$$

For each class we recorded the number of its members after discarding verbs whose frequency was less than 1 per 1M in the BNC. This gave us a first approximation of the size of each class. We then computed, for each polysemous verb, the total size of the classes of which it was a member. We calculated $p(class|amb\_class)$ by dividing the former by the latter (cf. equation (19)). We obtained an estimate for the class frequency $f(class)$ by multiplying $p(class|amb\_class)$ by the observed frequency of the verb in the BNC (cf. equation (18)).

[2]Our use of ambiguity classes is inspired by a similar use in HMM based part-of-speech tagging (Kupiec, 1992).

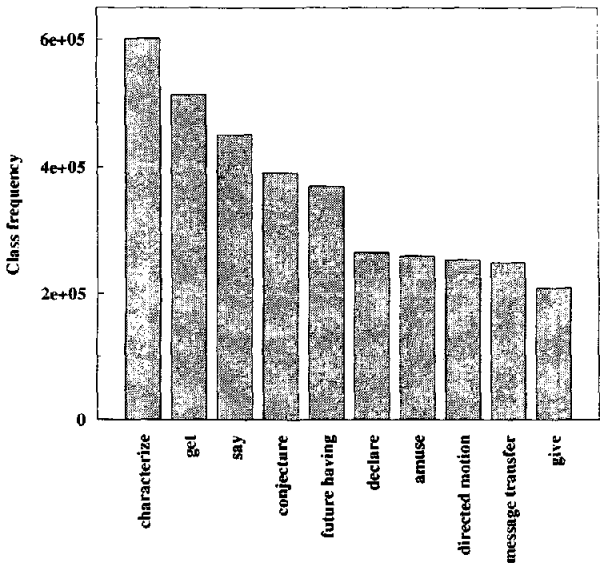

Figure 2: The ten most frequent classes

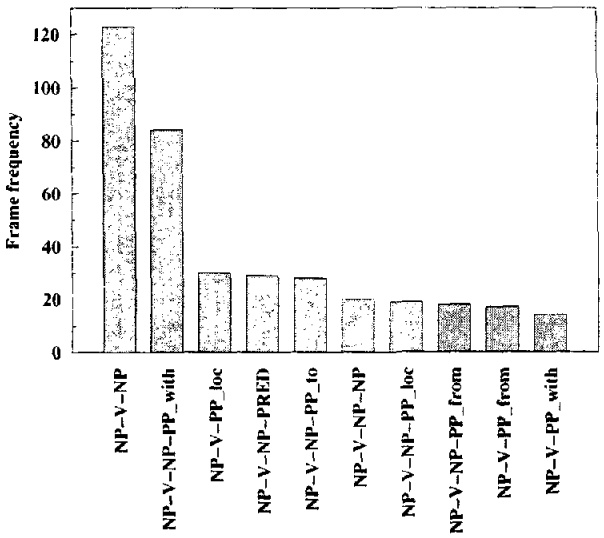

Figure 3: Ten most frequent frames in Levin

As an example consider the verb *pass* which has the classes THROW, SEND, GIVE and MARRY. The respective $p(class|amb\_class)$ for these classes are $\frac{27}{72}$, $\frac{20}{72}$, $\frac{15}{72}$ and $\frac{10}{72}$. By multiplying these by the frequency of *pass* in the BNC (19,559) we obtain the estimates for $f(verb, class)$ given in table 3. Note that simply relying on class size, without regard to frequency, would give quite different results. For example the class of MANNER OF SPEAKING verbs has 76 members, of which 30 have frequencies which are less than 1 per 1M, and is the seventh largest class in Levin's classification. According to our estimation scheme MANNER OF SPEAKING verbs are the 116th largest class. The estimates for the ten most frequent classes are shown in figure 2.

The estimation process described above involves at least one gross simplification, since $p(class|amb\_class)$ is calculated without reference to the identity of the verb in question. For any two verbs which fall into the same set of classes $p(class|amb\_class)$ will be the same, even though one or both may be atypical in its distribution across the classes. Furthermore, the estimation tends to favour large classes, again irrespectively of the identity of the verb in question. For example the verb *carry* has three classes, CARRY, FIT and COST. Intuitively speaking, the CARRY class is the most frequent (e.g., *Smoking can impair the blood which carries oxygen to the brain, I carry sugar lumps around with me*). However, since the FIT class (e.g., *Thameslink presently carries 20,000 passengers daily*) is larger than the CARRY class, it will be given a higher probability (0.45 versus 0.4). This is clearly wrong, but it is an empirical question how much it matters.

Finally, we wanted to estimate the probability of a given frame, $P(frame)$. We could have done this by acquiring Levin compatible subcategorization frames from the BNC. Techniques for the automatic acquisition of subcategorization dictionaries have been developed by Manning (1993), Briscoe and Carroll (1997) and Carroll and Rooth (1998). But the present study was less ambitious, and narrowly focused on the frames representing the dative and the benefactive alternation. In default of the more ambitious study, which we plan for the future, the estimation of $P(frame)$ was carried out on types and not on tokens. The mapping of Levin's linguistic specifications into surface syntactic information resulted in 79 different frame types. By counting the number of times a given frame is licensed by several semantic classes we get a distribution of frames, a sample of which is shown in figure 3.

The probabilities $P(frame|class)$ and $P(frame|verb)$ will be unreliable when the frequency estimates for $f(verb, frame)$ and $f(class, frame)$ are small, and ill-defined when the frequency estimates are zero. Following Hindle and Rooth (1993) we smooth the observed frequencies in the following way, where $f(V, frame) = \sum_i f(verb_i, frame)$, $f(V) =$

$\sum_i f(verb_i)$, $f(C,frame) = \sum_i f(class_i,frame)$ and $f(C) = \sum_i f(class_i)$. We redefine the probability estimates as follows:

$$(20)\ P(frame|verb) \approx \frac{f(verb,frame) + \frac{f(V,frame)}{f(V)}}{\sum_i f(verb,frame_i) + 1}$$

$$(21)\ P(frame|class) \approx \frac{f(class,frame) + \frac{f(C,frame)}{f(C)}}{\sum_i f(class,frame_i) + 1}$$

When $f(verb,frame)$ is zero, the estimate used is proportional to the average $\frac{f(V,frame)}{f(V)}$ across all verbs. Similarly, when $f(class,frame)$ is zero, our estimate is proportional to the average $\frac{f(C,frame)}{f(C)}$ across all classes. We don't claim that this scheme is perfect, but any deficiencies it may have are almost certainly masked by the effects of approximations and simplifications elsewhere in the system.

4 Results

We evaluated the performance of the model on all verbs listed in Levin which are polysemous and take frames characteristic for the dative and benefactive alternations. This resulted in 154 verbs which take the NP-V-NP-NP frame, 135 verbs which take the NP-V-NP-PP_{to} frame and 84 verbs which take the NP-V-NP-PP_{for} frame. The verbs were all polysemous and had an average of 3.8 classes. Each class had an average of 3.4 frames. Furthermore, we divided these verbs in two categories: verbs which can be disambiguated solely on the basis of their frame (e.g., *serve*; category A) and verbs which are genuinely ambiguous, i.e., they inhabit a single frame and yet can be members of more than one semantic class (e.g., *write*; category B).

The task was the following: given that we know the frame of a given verb can we predict its semantic class? In other words by varying the class in the term $P(verb,frame,class)$ we are trying to see whether the class which maximizes it is the one predicted by the lexical semantics and the argument structure of the verb in question.

For the verbs belonging to category A (306 in total) we used Levin's own classification in evaluation. The model's performance was considered correct if it agreed with Levin in assigning a verb the appropriate class given a particular frame. For class ambiguous verbs (category B) we compared the model's predictions against manually annotated data. Given the restriction that these verbs are semantically ambiguous in a specific syntactic frame we could not simply sample from the entire BNC, since this would decrease the chances of finding the verb in the frame we are interested in. Instead, for 31 class ambiguous verbs we randomly selected approximately 100 tokens from the data used for the acquisition of frame frequencies for the dative and benefactive alternation. Verbs with frame frequency less than 100 were not used in the evaluation.

The selected tokens were annotated with class information by two judges. The judges were given annotation guidelines but no prior training. We measured the judges' agreement on the annotation task using the Kappa coefficient (Siegel and Castellan, 1988) which is the ratio of the proportion of times, $P(A)$, that k raters agree to the proportion of times, $P(E)$, that we would expect the raters to agree by chance (cf. (22)). If there is a complete agreement among the raters, then $K = 1$, whereas if there is no agreement among the raters (other than the agreement which would be expected to occur by chance), then $K = 0$.

$$(22)\ K = \frac{P(A) - P(E)}{1 - P(E)}$$

We counted the performance of our model as correct if it agreed with the "most preferred", i.e., most frequent verb class as determined in the manually annotated corpus sample by taking the average of the responses of both judges.

We also compared the results for both categories to a naive baseline which relies only on class information and does not take subcategorization into account. For a given polysemous verb, the baseline was computed by defaulting to its most frequent class, where class frequency was determined by the estimation procedure described in the previous section.

As shown in table 4, in all cases our model outperforms the baseline. It achieves a combined precision of 91.8% for category A verbs. One might expect a precision of 100% since these verbs can be disambiguated solely on the basis of the frame. However, the performance of our model is less, mainly because of the way we estimated the terms $P(class)$ and $P(frame|class)$: we overemphasize the importance of frequent classes without taking into account how individual verbs distribute across classes.

The model achieves a combined precision of 83.9% for category B verbs (cf. table 4). Further-

	Category A			Category B		
Frame	Verbs	Baseline	Model	Verbs	Baseline	Model
NP-V-NP-NP	123	61.8%	87.8%	14	42.8%	85.7%
NP-V-NP-PP$_{to}$	113	67.2%	92%	15	73.4%	86.6%
NP-V-NP-PP$_{for}$	70	70%	98.5%	2	0%	50%
combined	306	65.7%	91.8%	31	61.3%	83.9%

Table 4: Model accuracy against baseline

Verb	Frame	Preferences
save	NP-V-NP-NP	GET, BILL
call	NP-V-NP-NP	GET, DUB
write	NP-V-NP-NP	MESSAGE TRANSFER, PERFORMANCE
make	NP-V-NP-NP	DUB, BUILD
extend	NP-V-NP-PP$_{to}$	FUTURE HAVING, CONTRIBUTE
present	NP-V-NP-PP$_{to}$	FULFILLING, REFLEXIVE APPEARANCE
take	NP-V-NP-PP$_{for}$	STEAL, PERFORMANCE
produce	NP-V-NP-PP$_{for}$	PERFORMANCE, CREATE

Table 5: Random sample of eight verbs and their semantic preferences as ranked by the model

more, our model makes interesting predictions with respect to the semantic preferences of a given verb. In table 5 we show the class preferences the model came up with for eight randomly selected verbs (class preferences are ranked from left to right, with the leftmost class being the most preferred one). Table 6 summarizes the average class frequencies for the same eight verbs as assigned to corpus tokens by the two judges together with inter-judge agreement (K). The category OTHER is reserved for corpus tokens which either have the wrong frame or for which the classes in question are not applicable. In general agreement on the class annotation task was good with Kappa values ranging from 0.68 to 1. As shown in table 6, with the exceptions of *call* and *produce* the model's predictions are borne out in corpus data.

Verb	Class			K
save	GET 64	BILL 25	OTHER 11	0.74
call	GET 2	DUB 94	OTHER 4	0.82
write	M. TRANS. 54	PERF. 19	OTHER 18	0.85
make	DUB 59	BUILD 20	OTHER 21	0.78
extend	FUT. HAV. 50	CONTR. 37	OTHER 13	0.71
present	FULFIL. 79	R. APP. 18	OTHER 3	0.94
take	PERF. 52	CREATE 13	OTHER 33	0.77
produce	PERF. 8	CREATE 91	OTHER 1	0.73

Table 6: Random sample of eight verbs and their semantic preferences as ranked by two judges

5 Discussion

This paper explores the degree to which syntactic frame information can be used to disambiguate verb semantic classes. In doing so, we cast the task of verb class disambiguation in a probabilistic framework which exploits Levin's semantic classification and frame frequencies acquired from the BNC. The approach is promising in that it achieves high precision with a simple model and can be easily extended to incorporate other sources of information which can influence the class selection process (i.e., selectional restrictions).

The semantic preferences which we generate can be thought of as default semantic knowledge, to be used in the absence of any explicit contextual or lexico-semantic information to the contrary (cf. table 5). Consider the verb *write* for example. The

model comes up with an intuitively reasonable ranking: we more often write things to people ("message transfer" reading) than for them ("performance" reading). However, faced with a sentence like *Max wrote Elisabeth a book* pragmatic knowledge forces us to prefer the "performance" reading versus the the "message transfer" reading. In other cases the model comes up with a counterintuitive ranking. For the verb *call*, for instance, the "get" reading (e.g., *I will call you a cab*) is preferred over the more natural "dub" reading (e.g., *John called me a fool*).

We still rely heavily on the verb class information provided by Levin. But part of original aim was to infer class information for verbs not listed by Levin. For such a verb, $P(class)$, and hence $P(verb, frame, class)$ will be zero, which is not what we want. Recent work in computational linguistics (e.g., Schütze (1993)) and cognitive psychology (e.g., Landauer and Dumais (1997)) has shown that large corpora implicitly contain semantic information, which can be extracted and manipulated in the form of co-occurrence vectors. The idea would be to compute the centroid (geometric mean) of the vectors of all members of a semantic class. Given an unknown verb (i.e., a verb not listed in Levin) we can decide its semantic class by comparing its semantic vector to the centroids of all semantic classes. We could (for example) determine class membership on the basis of the closest distance to the centroid representing a semantic class (cf. Patel et al. (1998) for a proposal similar in spirit). Once we have chosen a class for an unknown verb, we are entitled to assume that it will share the broad syntactic and semantic properties of that class.

We also intend to experiment with a full scale subcategorization dictionary acquired from the BNC. We believe this will address issues such as: (a) relations between frames and classes (what are the frames for which the semantic class is predicted most accurately) and (b) relations between verbs and classes (what are the verbs for which the semantic class is predicted most accurately). We also plan to experiment with different classification schemes for verb semantics such as WordNet (Miller et al., 1990) and intersective Levin classes (Dang et al., 1998).

References

Ted Briscoe and John Carroll. 1997. Automatic extraction of subcategorization from corpora. In *Proceedings of the 5th ACL Conference on Applied Natural Language Processing*, pages 356–363, Washinton, DC.

Lou Burnard, 1995. *Users Guide for the British National Corpus*. British National Corpus Consortium, Oxford University Computing Service.

Glenn Carroll and Mats Rooth. 1998. Valence induction with a head-lexicalized PCFG. In *Proceedings of the 3rd Conference on Empirical Methods in Natural Language Processing*, Granada.

Hoa Trang Dang, Karin Kipper, Martha Palmer, and Joseph Rosenzweig. 1998. Investigating regular sense extensions based on intersective Levin classes. In *Proceedings of the 17th International Conference on Computational Linguistics and 36th Annual Meeting of the Association for Computational Linguistics*, pages 293–299, Montréal.

A. P. Dempster, N. M. Laird, and D. B. Rubin. 1977. Maximum likelihood for incomplete data via the EM algorithm. *Journal of the Royal Statistical Society*, 39(2):1–38.

Donald Hindle and Mats Rooth. 1993. Structural ambiguity and lexical relations. *Computational Linguistics*, 19(1):103–120.

Frank Keller, Martin Corley, Steffan Corley, Matthew W. Crocker, and Shari Trewin. 1999. Gsearch: A tool for syntactic investigation of unparsed corpora. In *Proceedings of the EACL Workshop on Linguistically Interpreted Corpora*, Bergen.

Julian Kupiec. 1992. Robust oart-of-speech tagging using a hidden Markov model. *Computer Speech and Language*, 6(3):225–242.

Thomas K. Landauer and Susan T. Dumais. 1997. A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104(2):211–240.

Maria Lapata. 1999. Acquiring lexical generalizations from corpora: A case study for diathesis alternations. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*, College Park, MA.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago.

Christopher D. Manning. 1993. Automatic acquisition of a large subcategorization dictionary from corpora. In *Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, pages 235–242, Columbus, OH.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. 1990. Introduction to WordNet: An on-line lexical database. *International Journal of Lexicography*, 3(4):235–244.

Martha Palmer. 1999. Consistent criteria for sense distinctions. *Computers and the Humanities*, to appear.

Malti Patel, John A. Bullinaria, and Joseph P. Levy. 1998. Extracting semantic representations from large text corpora. In John A. Bullinaria, D. W. Glasspool, and G. Houghton, editors, *In Proceedings of the 4th Workshop on Neural Computation and Psychology*, pages 199–212. Springer, Berlin.

Adwait Ratnaparkhi. 1998. Unsupervised statistical models for prepositional phrase attachment. In *Proceedings of the 7th International Conference on Computational Linguistics*, pages 1079–1085.

Hinrich Schütze. 1993. Word space. In Stephen José Hanson, Jack D. Cowan, and C. Lee Giles, editors, *Advances in Neural Information Processing Systems*, pages 895–902. Morgan Kaufmann, San Mateo, CA.

Sidney Siegel and N. John Castellan. 1988. *Non Parametric Statistics for the Behavioral Sciences*. McGraw-Hill, New York.